B 12

Perception of voice quality in paralinguistic information types: A preliminary study

Caroline Menezes, Kikuo Maekawa (Nat. Inst. for Japanese Lang.) Hideki Kawahara (Wakayama Univ.)
{menezes, kikuo}@kokken.go.jp, kawahara@sys.wakayama-u.ac.jp

1. Introduction

In this paper we study the differences in voice quality due to variations in speaker attitudes. We treat speaker attitudes as paralinguistic information (PI) separate from active emotion following the distinction made by Fujisaki (1997). According to Fujisaki, paralinguistic information is conveyed by the speaker to the listener intentionally, while emotion is conveyed involuntarily. While paralanguage is distinct from spoken language (linguistic) it uses the same physical and psychological features as prosody such as pitch, rhythm, and loudness, besides voice quality (Laver, 1991, Maekawa and associates 1998, 2004, and 2000 to name a few).

Laver (1980) describes voice quality as the "auditory coloring of an individual speaker's voice" and acknowledges both laryngeal and supra-laryngeal effects on voice quality. Articulatory analyses of PI utterances by Maekawa (2004) support the influence of supra-laryngeal settings on voice quality. He reports a tendency for the forward displacement of the tongue dorsum for *suspicious* utterances when compared with those of *admiration*, which in contrasts shows a backward displacement compared. These results correlate strongly with raising and lowering of F2 (the second formant frequency) values (Maekawa, 2004, Maekawa & Kagomiya, 2000). The physiological displacement of the tongue is seen in the entire utterance including both consonants and vowels, leading to the speculation that this displacement is not just segmental manipulation, but prosodic manipulation.

In a follow-up study (same data) we report a clear differentiation in voice source characteristics (laryngeal) as a function of PI (Menezes & Maekawa, 2006). Of the four PI types studied, *i.e.*, *neutral, admiration, suspicion* and *disappointment*, it was found that the voice quality of *admiration* and *disappointment* was more breathy than *neutral* and *suspicion*. Spectral analysis revealed slow (degree of spectral slope) and incomplete closure of the glottis (difference between first and second harmonic) for *admiration* and *disappointment*. In contrast, in *suspicion* utterances we found a quick closing phase and complete closure of the glottis. Since clear differences in voice quality exist for different PI attitudes is it possible that people use them to distinguish one PI from another? Different PI types can be produced with different voice quality but there is no one to one relationship between them, for example, *admiration* and *disappointment* are similar on one hand, but different from *neutral* and *suspicion*. So we can assume that in the absence of other important PI cues like F0 contour and duration two PIs sharing similar voice characteristics should be more difficult to discriminate than two PIs with contrasting voice characteristics. In this preliminary study we attempt

to answer these questions by conducting a perceptual experiment using PI utterances that have been normalized for pitch and duration but conserve voice quality (both laryngeal and supra-laryngeal) differences.

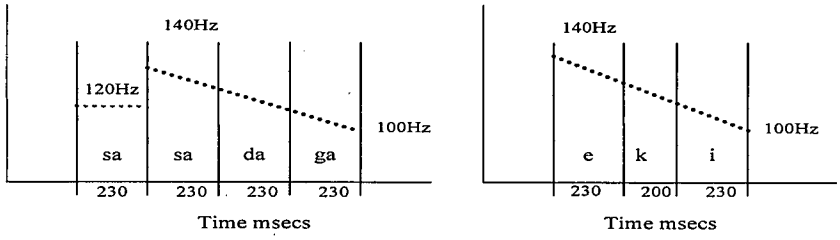

Figure1: Schematic representation of morphed 'sasadaga' (left) and 'e¯ki' (right). Solid lines demarcate moras and dotted line show stylized F0 contour.

2. Method

2.1 Stimuli Morphing

The data analyzed consists of two native Japanese speaker's production of the short phrase /sasadaga/ (surname 'Sasada' followed by nominal case particle 'ga') and /eki/ (station) with four different PI types namely, *neutral*, *admiration*, *suspicion* and *disappointment*. The data were recorded using the EMA (Carstens 100) system (NTT basic research laboratory, Atsugi, Japan). In this study, however, only the acoustic signals are analyzed. To create the stimuli for the perception test, five utterances per attitude for each speaker were morphed for F0 and duration while leaving unchanged voice quality differences. Morphing was done using the program STRAIGHT developed specifically as a tool for carrying out speech perception tests on natural speech stimuli (Kawahara et al., 1999, Kawahara (submitted), and Matsui & Kawahara 2003). STRAIGHT uses three real valued parameters (namely, STRAIGHT spectrogram, an aperiodicity map and F0 with voicing information) that allows for independent manipulation of parameters without causing interaction between the manipulated values (Kawahara (submitted)). The stimuli were morphed in such a way that the duration of every mora in 'sasadaga' was set to 230ms length regardless of PI type. This duration was determined by finding the average of all moras in a *neutral* utterance. 'Eki' was treated differently. For 'eki' to sound natural it was found that /k/ had to be rather long. For this reason the vowels /e and i/ were set to 230ms in duration (average duration for /e & i/ in *neutral* utterance) and /k/ was maintained at 200ms. In determining the appropriate duration value various methods were attempted but the one described above was found to produce the most natural sounding stimuli. Once the stimuli were morphed for uniform duration they were then morphed for F0. 'Sasadaga,' a phrase without lexical pitch accent, is produced in *neutral* utterances with a low pitch on the first mora and a rise in F0 in the second mora and a gradual declination towards the end of the phrase. In the stylized F0 pattern we maintain this *neutral* pattern by setting the F0 in the first mora at 120Hz. At

the start of the second mora the F0 is set at 140 Hz which gradually declines to 100Hz at the end of the utterance. This pattern was maintained for all PI types for the utterance 'sasadaga'. However, 'eki' had to be treated differently since it has a lexical pitch accent on the first mora. The stylized F0 contour for 'eki' was created by starting the utterance with a F0 value at 140 Hz with a steep decline to 100Hz by the end of the utterance. See figure 1 for a schematic representation of the morphed stimuli.

2.2 Perception Experiment

The stimuli were then presented in a forced choice perception experiment, where subjects listened to each stimulus and judged if they were *neutral*, *admiration*, *suspicion* or *disappointment* on a computer screen that gave them all four options. The experiment consisted of two sessions separated by speaker (A & B). Each speaker's session consisted of a practice and a test. For each practice session, a single example of 'eki' and 'sasadaga' for each PI type was selected and presented twice (1x4x2x2=16). Each test session consisted of 5 productions of each PI type for both 'sasadaga' and 'eki' repeated 5 times (5x4x2x5=200). Stimuli were randomized and half the listeners were presented speaker A first then speaker B, and the other half were presented speaker B first then speaker A. In the instructions subjects were able to hear single examples of the unmodified utterances for each PI type, and they were informed that the experiment items consisted of computer modified versions of similar utterances They were also instructed that most of the obvious cues of speaker attitude had been modified and they would have to make their judgment based on the residual nuances. 20 subjects participated in the perception experiment, 8 males (avg. age 29) and 12 females (avg. age 37).

3. Results

In tables 1 (for sasadaga) and 2 (for eki), we tabulate the responses of males and females separately. The rows list the different PIs analyzed in this experiment and the columns give subject responses. The bold numbers on the diagonal indicate the number of stimuli that were correctly judged for PI type, and non-diagonal numbers give the number of error judgments (*i.e.*, responses did not agree with stimuli in PI type). Bracketed numbers give the percentage (for correct responses only). Percentage values greater than 25 indicate that subjects are able to perceive the correct PI type above chance level in a four choice task.

From the 20 subjects, two (one male and one female) were considered as outliers as they contained unusually high number of error responses. The percentage values in tables 1 and 2 show that subjects were able to distinguish well above chance the correct PI type based on differences in voice quality alone (except for 'eki' *admiration*: males were around chance level 25% and females well below). Comparing the percentage of correct responses between the two tables (1 & 2) we see that subjects were better in judging PI for 'sasadaga' than for 'eki'. In general, subjects found no difficulty in determining *neutral* PI type (96% for 'sasadaga' and 50% for 'eki'). In the case of 'eki' it may

appear that *disappointment* was best perceived of all PI types, but we will explain later reasons why we should be cautious in interpreting too much into this trend.

For non-*neutral* utterances subjects were best at perceiving *suspicion*. Further if you examine the

Table 1: Response against actual PI types for 'sasadaga' . Bold numbers along the diagonal give sum of correct responses (percentage), a = *admiration*, d = *disappointment*, n = *neutral*, and s = *suspicion*.

		Response				
Gender	PI_Type	a	d	n	s	Total
f	a	**310 (56%)**	66	96	78	550
	d	130	**285 (52%)**	50	85	550
	n	21	4	**518 (94 %)**	7	550
	s	67	3	79	**401 (73%)**	550
f Total		528	358	743	571	2200
m	a	**157 (45%)**	71	65	57	350
	d	82	**154 (44%)**	71	43	350
	n	1		**343 (98%)**	6	350
	s	84	10	98	**158 (45%)**	350
m Total		324	235	577	264	1400
Total		852	593	1320	835	3600

errors made for this PI type (i.e., compare across row 's' for PI type in both tables) you notice that when it was not judged correctly it was most often and consistently by all speakers mistaken for *neutral*. The other two PI types namely, *admiration* and *disappointment* proved to be more difficult for the subjects to differentiate. Almost always erroneous responses of *disappointment* were confused for *admiration* by all speakers. In contrast, incorrect responses for *admiration* were judged to be *disappointment* except for females in the case of 'sasadaga'. However, note that for 'eki' subjects were poor in correctly perceiving admiration (females were even below chance level). They perceived it overwhelmingly to be disappointment. In fact, after the test most subjects even commented that they choose disappointment instead of admiration because of the falling intonation contour.

Table 2: Response against actual PI types for 'eki'. Bold numbers along the diagonal give sum of correct responses (percentage), a = *admiration*, d = *disappointment*, n = *neutral*, and s = *suspicion*.

		Response				
Gender	piType	a	d	n	s	Total
f	a	**86 (16%)**	373	75	16	550
	d	51	**430 (78%)**	52	17	550
	n	43	150	**281(51%)**	76	550
	s	57	69	181	**243 (44%)**	550
f Total		237	1022	589	352	2200
m	a	**102 (27%)**	197	49	27	375
	d	82	**228 (61%)**	45	20	375
	n	97	57	**185 (49%)**	36	375
	s	85	29	130	**131 (35%)**	375
m Total		366	511	409	214	1500
Total		603	1533	998	566	3700

4. Discussion

These results agree with the assumptions made in the introduction. First, we asked the question, can people distinguish different PI types based only on voice quality. Note that in our morphed stimuli we manipulated F0 contour and duration only, so voice quality here includes both supra-laryngeal differences as noted by Maekawa and associates (Maekawa, 2004 and Maekawa & Kagomiya, 2000) and laryngeal characteristics (Menezes & Maekawa, 2006). The results of this study indicate that people do and will use voice quality to distinguish differences in PI types. Though correct responses were above chance they were not very high. This makes us believe that voice quality, however, may not be the most important cue, and at least for the case of Japanese, intonation changes might be the best indication of PI type. Note the high correct perception of *neutral* utterances for 'sasadaga' but not for 'eki'. If we go back to figure 1, we see that in the case of 'sasadaga' the *neutral* intonation pattern was used for the morphing of all PI types, but this was not the case for 'eki'. For 'eki' the lexical pitch accent is on the first mora and the peak occurs later in the mora but in our F0 stylization this peak occurs at the very beginning of the mora and it falls from there on. One more result to support this assumption is the predominance of *disappointment* responses for 'eki'. Given the specific difficulty to differentiate the voice quality in *admiration* and *disappointment* for the utterance 'eki', the steadily falling contour biased subject's responses to *disappointment*. In naturally occurring *admiration* utterances there is a steep F0 rise and fall within the utterance.

In our earlier study (Menezes & Meakawa 2006) we reported that *admiration* and *disappointment* shared similar spectral characteristics, which were different from that of *neutral* and *suspicion*. Based on these findings we postulated that if people can indeed differentiate PI types from voice quality, then in the absence of other cues, PI types that have similar voice source characteristics should be more difficult to differentiate than PI types with dissimilar characteristics. Again our results seem to validate this postulation. Subjects found it easier to differentiate *neutral* and *suspicion* but not *admiration* and *disappointment*. Two striking results back this claim; first the higher percentage of correct responses for *neutral* and *suspicion* when compared to *admiration* and *disappointment*. And the second lies in the error judgments. Subjects confused *admiration* for *disappointment* and *disappointment* for *admiration* more than they confused the other PI types. Moreover, in the earlier study (Menezes & Maekawa, 2006) we showed that *neutral* and *suspicion* were different from *admiration* and *disappointment* but they were speaker differences and strong syllable effects, for example the first mora in 'sasadaga' in the case of *suspicion* is produced with a creaky voice due to the accompanying very low pitch but this is not seen in *neutral* utterances.

5. Conclusion

This preliminary study has shown that the program STRAIGHT can be effectively used for studying the perception of voice quality in different attitudes. The drastic manipulation of duration and F0 still resulted in natural sounding stimuli from which people were able to perceive differences in voice

quality.

While voice quality might not be the main cue, we suppose that voice quality helps in making quantitative estimation of speakers' attitude, for example, how disappointed or how suspicious the speaker is. However, further experiments need to be conducted using stylized but attitude specific F0 contours to show the relative importance of source spectrum, F0 contour and duration in attitude or paralinguistic discrimination.

Acknowledgment

The authors would like to thank Osamu Fujimura, Yosuke Igarashi and Donna Erickson for many useful suggestions on this topic and paper. Thanks to Atsuko Asami for strategic help with the experiment.

References

Fujisaki, H. (1997) Prosody, models, and spontaneous Speech. In Computing Prosody, Sagisaka et al. (ed.), Springer, 27-40.

Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A. (1999) Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction. Speech Communication, Vol. 27, No. 3-4, pp. 187-207.

Kawahara, H (submitted) STRAIGHT: Exploitation of the other aspect of VOCODER. Perceptually isomorphic decomposition of speech sounds. In Acoustic Science and Techonology 2006.

Matsui, H. and Kawahara H. (2003) Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system. In Eurospeech'03, pp. 2113-2116, Geneva.

Laver, J. (1980) The phonetic descriptionof voice quality. Cambridge University Press.

Laver, J. (1991) The gift of speech. Edinburgh University Press.

Maekawa, K. (2004) Production and perception of 'Paralinguistic' information. In Proceedings of Speech Prosody, Nara, Japan, 367-374.

Maekawa, K. and Kagomiya, T. (2000) Influence of paralinguistic information on segmental articulation. In 6th International Conference on Spoken Language Processing, Beijing, China, v2, 349-52.

Maekawa, K. (1998) Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. In 5th International Conference on Spoken Language Processing, Sydney, Australia, v2, 635-38.

Menezes, C. and Maekawa, K. (2006) Paralinguistic effects on voice quality: A study in Japanese. In Proceedings of Speech Prosody, Dresden, Germany.

Takahashi, T., Fujii, T., Nishi, M., Banno, H. Irino, T. and Kawahara, H. (2005). Voice and Emotional Expression Transformation based on Statistics of Vowel Parameters in an Emotional Speech Database. In Interspeech2005, pp.1853-1856, Lisboa.